Bayesian Networks

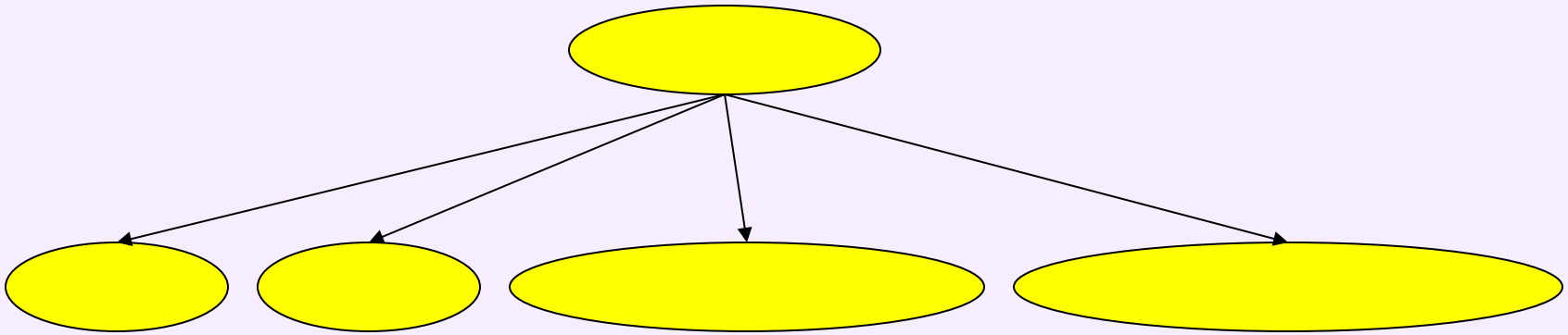
Master: Modelling for Science and Engineering

Module: DATA VISUALIZATION AND MODELLING

Rosario Delgado

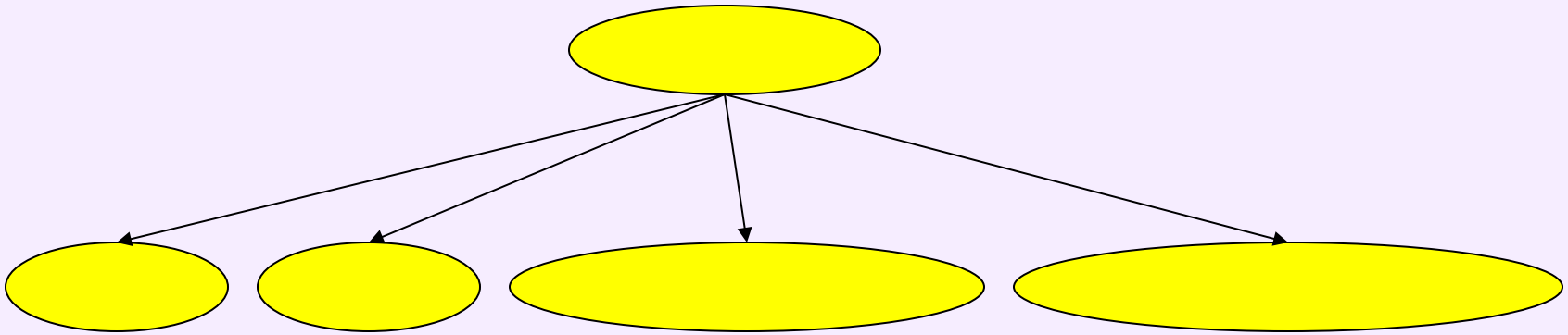
Department of Mathematics

University Autonomous of Barcelona



- Knowledge acquisition (*learning*) is one of the bottlenecks in expert systems building.

- Bayesian Networks (BN) are the soundest framework for uncertain reasoning and offer an important advantage for model construction with respect to other expert systems: we construct the most probable BN given the observed cases (*learning*), and this model provides the optimal prediction for future cases (*inference*). **White-box!**



- BN are graphical structures for representing the probabilistic relationships among a large number of variables and for doing probabilistic inference with those variables.
- In the opinion of many AI researchers, BN are the most significant contribution in AI in the last years.
- Many applications: spam filtering, speech recognition, robotics, diagnostic systems, health care, risk assessment, …

Block 1: Basics

1. An introductory example.
2. Conditional independence.
3. The Markov condition.
4. Bayesian networks.
5. Diagnosis and Prognosis with BN.
6. Causality.
7. Causal networks.
8. References.

1. An introductory example

Sam takes the ELISA test and it comes back "+" for HIV.

How likely is it that Sam is infected with HIV?

To answer we need to know the accuracy of the ELISA test. On these tests, we usually know:

✓the **true positive rate** (*sensitivity*)

P(ELISA= + / HIV=present)=0.999

✓ the **true negative rate** (*specificity*)

P(ELISA= − / HIV=absent)=0.998

The probability of Sam being infected with HIV knowing the test is "+" is

$$P(HIV=present \,/\, ELISA=+) = ?$$

We can compute this probability by using **Bayes' Theorem** if we know

$$P(HIV = present)=0.00001$$

Then, by **Bayes' Theorem**:

$$P(HIV = present \,/\, ELISA = +) = \frac{P(ELISA = + \,/\, HIV = present)P(HIV = present)}{P(ELISA = +)}$$

In order to compute the probability in the denominator, we use the **Law of Total Probability**:

$$P(ELISA=+)=P(ELISA=+/HIV=present)P(HIV=present)$$
$$+P(ELISA=+/HIV=absent)P(HIV=absent)$$
$$=0.999\times 0.00001+(1-0.998)\times(1-0.00001)=0.00200997$$

Then, by the Bayes' Theorem:

$$P(HIV=present/ELISA=+)$$
$$=\frac{P(ELISA=+/HIV=present)P(HIV=present)}{P(ELISA=+)}$$
$$=\frac{0.999\times 0.00001}{0.00200997}=0.004970223436$$

$$P(HIV=present/ELISA=+)\cong 0.00497$$

5 people has HIV present out of any 1000 to which ELISA= +

- P(HIV=present) is the **prior probability**:

  the probability of the event **before** updating using new information.

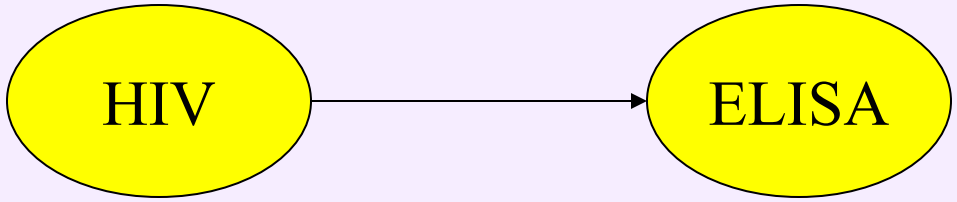
  In this case, P(HIV=present) =0.00001.

- P(HIV=present / ELISA= +) is the **posterior probability**:

the probability of the event **after** its prior probability has been updated based on new information ("**evidence**").

In this case, P(HIV=present / ELISA= +) ≅ 0.005.

We can represent the knowledge in this example by a **two-node Bayesian network:**



P(HIV=present)=0.00001

P(HIV=absent)=0.99999

(**prior** probabilities)

P(ELISA=+ / HIV=present)=0.999

P(ELISA= − / HIV=present)=0.001

P(ELISA=+ / HIV=absent)=0.002

P(ELISA= − / HIV=absent)=0.998

(**conditional** probabilities)

From **prior** probabilities and **conditional** probabilities we obtained **posterior** probabilities by using **Bayes' Theorem**.

$$P(HIV = present \,/\, ELISA = +) \cong 0.00497$$

2. Conditional Independence

(Ω,P) probability space. **A**, **B** and **C** sets containing discrete random variables defined in Ω.

The sets **A** and **B** are said to be **conditionally independent given the set C** if for any variables in the sets, say A, B and C, respectively, and for any values these variable can take, say a, b and c, whenever $P(C=c) \neq 0$, events $A=a$ and $B=b$ are independent given event $C=c$. That is:

$$P(A=a,\ B=b \,/\, C=c) = P(A=a/\ C=c)\ P(B=b \,/\, C=c)$$

Equiv.: $P(A=a/\ B=b,\ C=c) = P(A=a/\ C=c)$ if $P(B=b) \neq 0$

$$I_P(A,B/C)$$

Knowing C tells everything about A and no gain by knowing B (either because B does not influence A, or because knowing C provides all information knowing B would give).

An example

You roll a blue dice and a red dice.
The two results are independent of each other.
Let A="result of the blue dice"
B="result of the red dice".
Then, A and B are independent.
Let C="the sum of results".
Consider a=1, b=3, c=4. Then,

$P(A=a/\ B=b,\ C=c) = 1$, while $P(A=a/\ C=c)=P(B=b)=1/6 \neq 1$

That is, **although A and B are independent, they are not conditionally independent given C**.

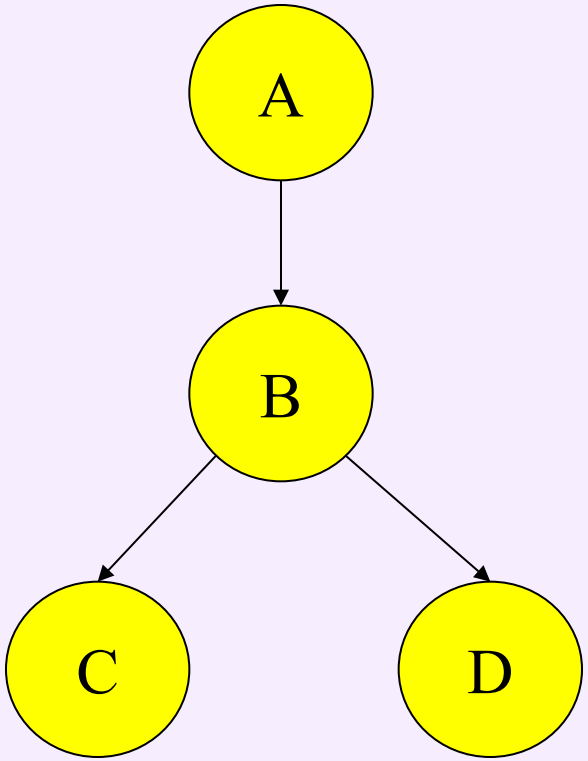
3. The Markov condition

A **directed graph** is a pair (V, E), where V is a finite set whose elements are called **nodes**, and E is a set of ordered pairs of nodes called **edges**, **arcs** or **arrows**.

Example:

V=$\{A,B,C,D\}$

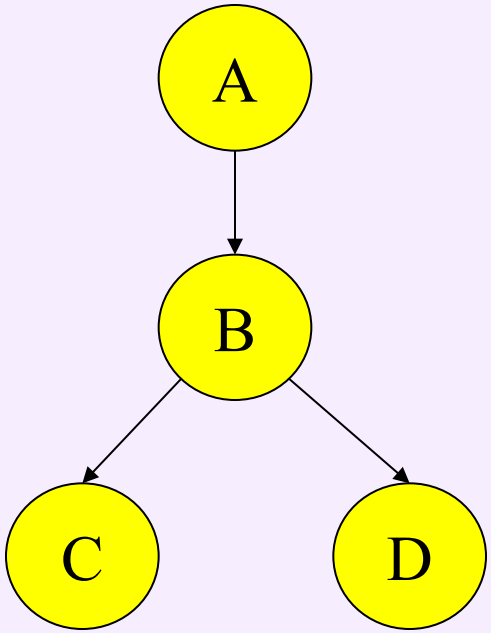
E=$\{(A,B), (B,C), (B,D)\}$



A **path** from node *X* to node *Y* is a set of arrows connecting them. Example: a path from *A* to *D* is $\{(A,B), (B,D)\}$.

A **directed cycle** is a path from a node to itself.

A **directed graph** (V,E) is called a **directed acyclic graph (DAG)** if it contains no cycles.



Given a DAG, a node *X* is a **"parent"** of a node *Y* if there is an arrow from *X* to *Y* (example: *A* is a parent of *B*, *B* is a parent of *C* and *D*). Then, *Y* is a **"descendant"** of *X*.

A is an ***"ancestor"*** of *C* (path from *A* to *C*).

Node *A* is a **"root"** since it has no parents.
Nodes *C* and *D* are **"leaf"** since they have not children.
B is a "intermediate node".

DEFINITION:

Let P be a prob. distribution of the r.v. in a set V, and let Γ=(V, E) be a DAG.

We say that (Γ, P) satisfies the **Markov condition** if for each *X* in V,

{X} is conditionally independent of the set of all its non-descendants given the set of all its parents:

$$I_P(\{X\}, ND(X) / PA(X))$$

(*ND(X)* denote the non-descendants of *X*, and *PA(X)* its parents)

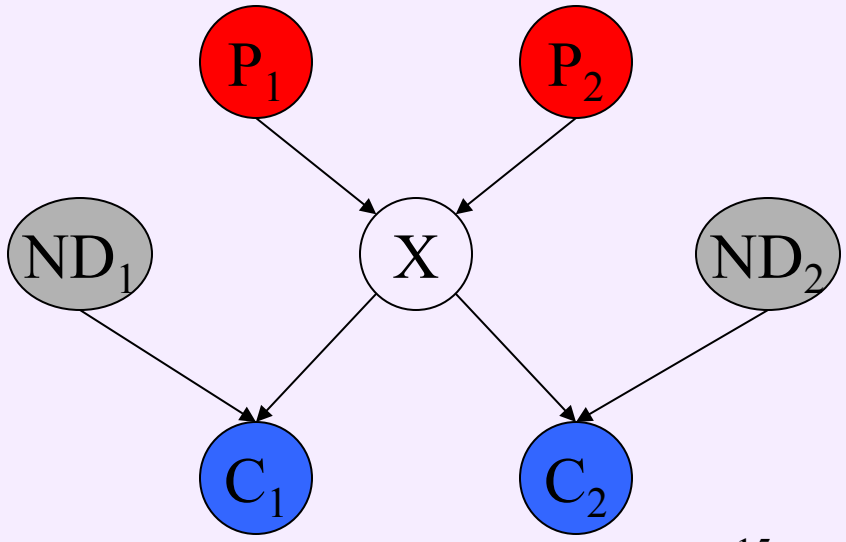
Note: If X is a root, PA(X)=empty: {X} and ND(X) are independent.

As PA(X) is included in ND(X), Markov condition is equivalent to:

$$I_P(\{X\}, ND(X) - PA(X) / PA(X))$$

In the example, *X* is **conditionally independent** of each of its non-descendants that are not parents of *X*, ND_1 and ND_2, given the set of its parents $\{P_1, P_2\}$:

It is by using this expression that we check the Markov cond. relative to a particular (Γ, P) !



THEOREM 1: (Theorem 1.4, Neapolitan, 2004)
If (Γ, P) satisfies the Markov condition, then P is equal to the product of the conditional distributions of all nodes given the values of their parents, whenever these conditional distributions exist.

That is, if V={X1,…,Xn}, for all possible values xi of Xi, we have

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i \,/\, PA(X_i))$$

(the **"Chain Rule"**)

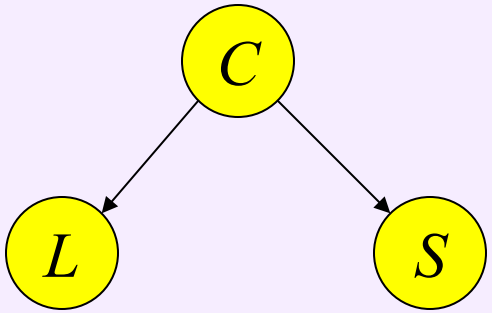
Idea: If we start with a joint distribution P satisfying the Markov condition with some DAG Γ, the probabilities of P will be given by the product of the conditional distributions.

Example 1

Let Ω be the 13 objects, P assigns 1/13 to each one.

Variable	Value	Outcome
L *(Letter)*	ℓ_1	Objects with A
	ℓ_2	Objects with B
S *(Square)*	s_1	Square objects
	s_2	Circular objects
C *(Color)*	c_1	Black objects
	c_2	White objects

V={L,S,C}
E={(C,L), (C,S)}
Γ=(V, E) is a DAG



Node	Parents	Non-desc.	Cond. Independence
L	*C*	*S*	*L* and *S*, given *C*
S	*C*	*L*	idem
C	Null set	Null set	None

Then, (Γ, P) verifies the Markov condition if *L* and *S* are conditionally independent given *C*:

$$P(L, S / C) = P(L / C)\, P(S / C)$$

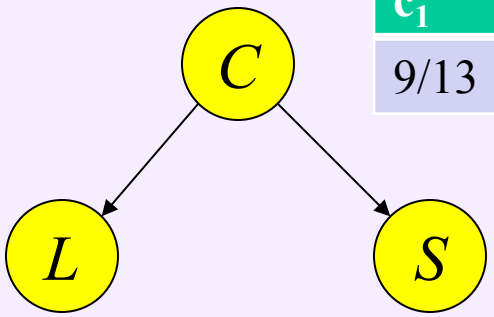
Indeed, $P(L, S / C) = P(L / C)\, P(S / C)$ since

If $C = c_1$	ℓ_1	ℓ_2	
s_1	2/9	4/9	2/3
s_2	1/9	2/9	1/3
	1/3	2/3	

If $C = c_2$	ℓ_1	ℓ_2	
s_1	1/4	1/4	1/2
s_2	1/4	1/4	1/2
	1/2	1/2	

As (Γ, P) verifies the Markov condition, by Theorem 1, in order to find any value of the joint distribution P, we only need to determine the conditional distributions of the variables for that DAG.

Conditional distributions (CPTs) of L and S with respect to C:



c_1	c_2
9/13	4/13

Marginal distribution of the root C.

L	ℓ_1	ℓ_2
If C = c_1	1/3	2/3
If C = c_2	1/2	1/2

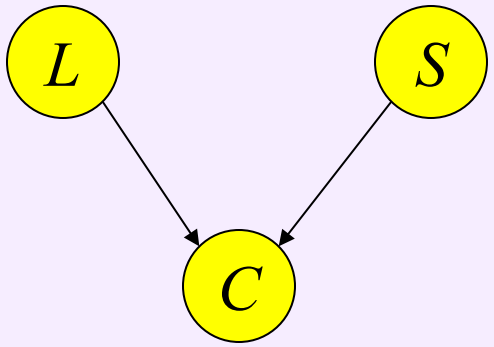
S	s_1	s_2
If C = c_1	2/3	1/3
If C = c_2	1/2	1/2

Joint probability distribution P:
2^3-1=8-1=7 values.
CPTs: 1+2+2=5 values.

We know: $P(C=c_1, L=\ell_1, S=s_1)=2/13$

But we could use Theorem 1 (the "Chain Rule") to compute this joint probability in this way: $P(C=c_1, L=\ell_1, S=s_1)=$
$P(C=c_1)\ P(L=\ell_1 / C=c_1)\ P(S=s_1 / C=c_1)=(9/13)(1/3)(2/3)=2/13$
(analogously for the other 7 probabilities).

Not in any DAG the Markov condition holds!



V={L,S,C}, E={(L,C), (S,C)}
Γ=(V, E) is a DAG

Node	Parents	Non-desc.	Cond. Independence
L	Null set	*S*	*L* and *S* (independent)
S	Null set	*L*	idem
C	*L, S*	Null set	None

For P to satisfy the Markov condition with DAG Γ we must have *L* and *S* independent. Are they?... Not really:

$P(L=\ell_1, S=s_1)=3/13$

$P(L=\ell_1)\ P(S=s_1)=(5/13)(8/13)=(3/13)(40/39) \neq 3/13$

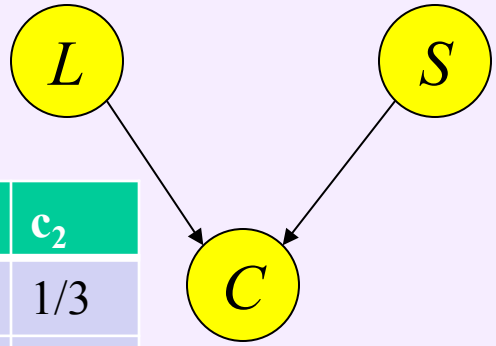
so, *L* and *S* are not independent!

(Γ, P) does not satisfy the Markov condition.

The **"Chain Rule"** also fails. Indeed,

ℓ_1	ℓ_2
5/13	8/13

s_1	s_2
8/13	5/13



C	c_1	c_2
If L=ℓ_1, S=s_1	2/3	1/3
If L=ℓ_1, S=s_2	1/2	1/2
If L=ℓ_2, S=s_1	4/5	1/5
If L=ℓ_2, S=s_2	2/3	1/3

$P(C=c_1, L=\ell_1, S=s_1) = 2/13$

$P(L=\ell_1)\, P(S=s_1)\, P(C=c_1 / L=\ell_1, S=s_1) = (5/13)(8/13)(2/3) = (2/13)(40/39)$

$\neq 2/13$

The chain rule failure implies the Markov condition fails!

Recall Theorem 1: If we start with a joint distribution P on a sample space Ω satisfying the **Markov condition** with some DAG Γ, the probabilities of P are the product of the corresponding conditional distributions (**"Chain Rule"**).

In practice, we do not specify (Ω, P) from which compute the conditional distributions. Rather, we identify r.v. and their conditional distributions directly. Is the product of these conditional distributions, then, a joint distribution satisfying the Markov condition?

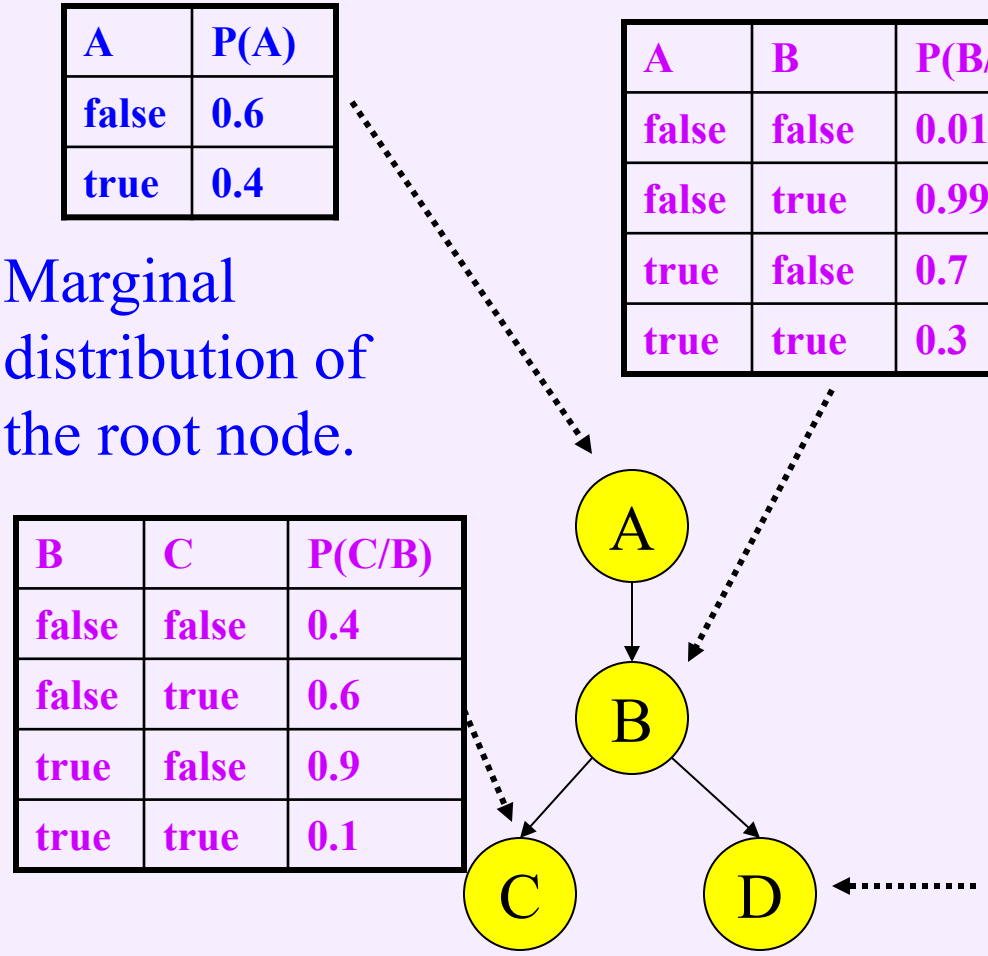
THEOREM 2: (Theorem 1.5, Neapolitan, 2004)
Let Γ be a DAG in which each node is a discrete r.v., and let a conditional distribution of each node given the values of its parents in Γ be specified. Then, the product of these conditional distributions yields a joint probability distribution P of the variables, and (Γ, P) satisfies the Markov condition.

Example 2

A	P(A)
false	0.6
true	0.4

Marginal distribution of the root node.

A	B	P(B/A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3



B	C	P(C/B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

B	D	P(D/B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

The conditional probability distribution of each non-root node X with respect to its parents $P(X_i / Parents(X_i))$ quantifies the effect of the parents on the node (CPTs).

The parameters are the probabilities in these tables.

Total parameters: 1+2+2+2=7.

Conditional Probability Distribution of *C* with respect to its parent *B*:

B	C	P(C/B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

For a given combination of values of the parents (called "instantiation"), *B* in this example, we have:

P(C=true / B=false) + P(C=false / B=false)=1

P(C=true / B=true) + P(C=false / B=true)=1

Theorems 1 and 2 enable to reduce the problem of determining a huge number of probability values to that of determine relatively few:

For a Boolean variable (2 values: True/False) with k Boolean parents, its table has $\mathbf{2^{k+1}}$ probabilities (but only $\mathbf{2^k}$ need to be stored). If k is not large, it is a manageable number. If the DAG contains n nodes, the total is $\mathbf{\leq 2^k\ n}$.

Instead, we would need $\mathbf{2^n - 1}$ values for the joint distribution P, which usually is a far greater number!

The joint probability distribution

If we want to calculate the joint probability of the four variables, we use the ***chain rule***:

P(A = true, B = true, C = true, D = true) = ?

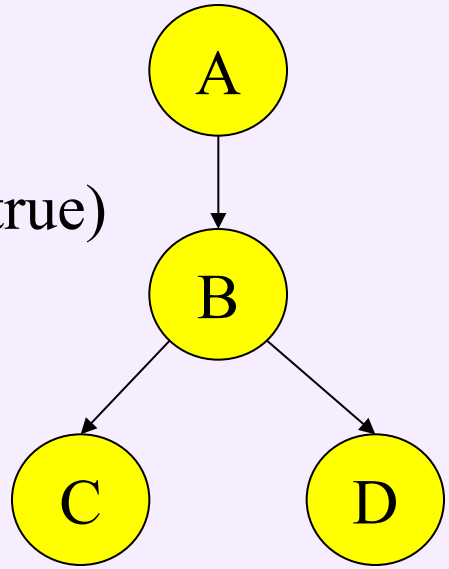
P(A = true, B = true, C = true, D = true)

From the graph structure:

= P(A = true) P(B = true / A = true)
P(C = true / B = true) P(D = true / B = true)

= 0.4 × 0.3 × 0.1 × 0.95 = **0.0114**

From the CPTs and the marginal distribution for the root A



And analogously for the other $2^4 - 1 = 15$ probabilities.

4. Bayesian Networks

DEFINITION:

Let P be a joint probability distribution of the r.v. in a set V, and let $\Gamma=(V, E)$ be a DAG.

We call (Γ, P) a **Bayesian network (BN)** if (Γ, P) satisfies the **Markov condition.**

Owing to Theorem 1, P is the product of its conditional distributions in Γ, and this is the way P is always represented in a BN.

Furthermore, owing to Theorem 2, if we specify a DAG and any discrete conditional distributions, we obtain a Bayesian network.

This is the way in which BN are constructed in practice!

Example 3

Ω={athletes} (#Ω=N), P assigns 1/N to each athlete. Let three discrete r.v. on (Ω, P):

X_1 = athlete is doped (T/F)

X_2 = doping test A (blood test) is positive (T/F)

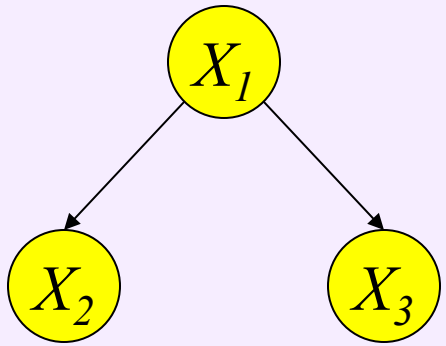
X_3 = doping test B (urine test) is positive (T/F)

- ✓ The two tests are **not** independent: if test A is +, then test B will "probably" give +.
- ✓ But they are independent if we know the result of X_1: $P(X_3 / X_1, X_2) = P(X_3 / X_1)$

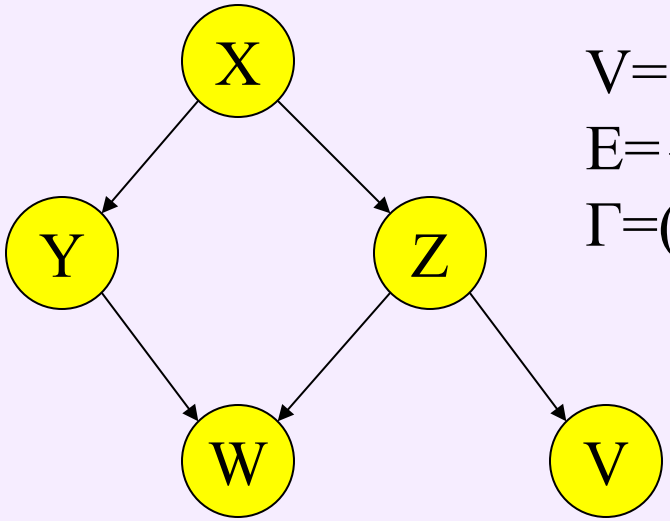
  $P(X_2 / X_1, X_3) = P(X_2 / X_1)$
- ✓ As a consequence, the joint probability distribution can be calculated as:

$$P(X_1, X_2, X_3) = P(X_1)P(X_2 / X_1)P(X_3 / X_1)$$

By Theorem 2 the following DAG and P form a BN that models the influences among X_1, X_2 and X_3:



Example 4

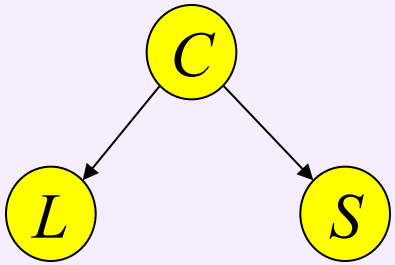


V={X, Y, Z, W, V},
E={(X,Y), (X,Z), (Y,W), (Z,W), (Z,V)}
Γ=(V, E) is a DAG

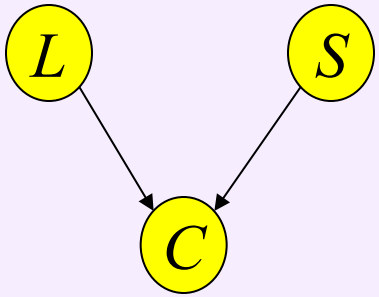
If a probability P with this DAG satisfies the Markov condition ((Γ,P) is a BN), we would have the following conditional independencies:

Node	Parents	Nondescendents	Conditional Independency
X	Null set	Null set	None
Y	*X*	*Z, V*	*Y* and {*Z, V*}, given *X*
Z	*X*	*Y*	*Z* and *Y*, given *X*
W	*Y, Z*	*X, V*	*W* and {*X, V*}, given {*Y, Z*}
V	*Z*	*X, Y, W*	*V* and {*X, Y, W*}, given *Z*

Coming back to Example 1



V={L,S,C}, E={(C,L), (C,S)}, Γ=(V, E) is a DAG.
(Γ, P) is a **Bayesian Network** since verifies the Markov condition (*L* and *S* are conditionally independent given *C*). By Theorem 2, the conditional distributions (CPTs) and the marginal distribution of the root nodes, give the joint probability P by Chain Rule.



V={L,S,C}, E={(L,C), (S,C)}, Γ=(V, E) is a DAG.
(Γ, P) is **NOT a Bayesian Network** since P does not satisfy the Markov condition with this DAG (*L* and *S* are **not** independent). There is **no** reason to suspect P would be the product of the cond. dist. in the DAG. Indeed, we saw that it is not!! (Chain Rule fails)

Previous example illustrates that, if we develop a BN from an arbitrary DAG Γ and the conditionals of a probability distribution P relative to that DAG, in general the resultant Bayesian network does not contain P.

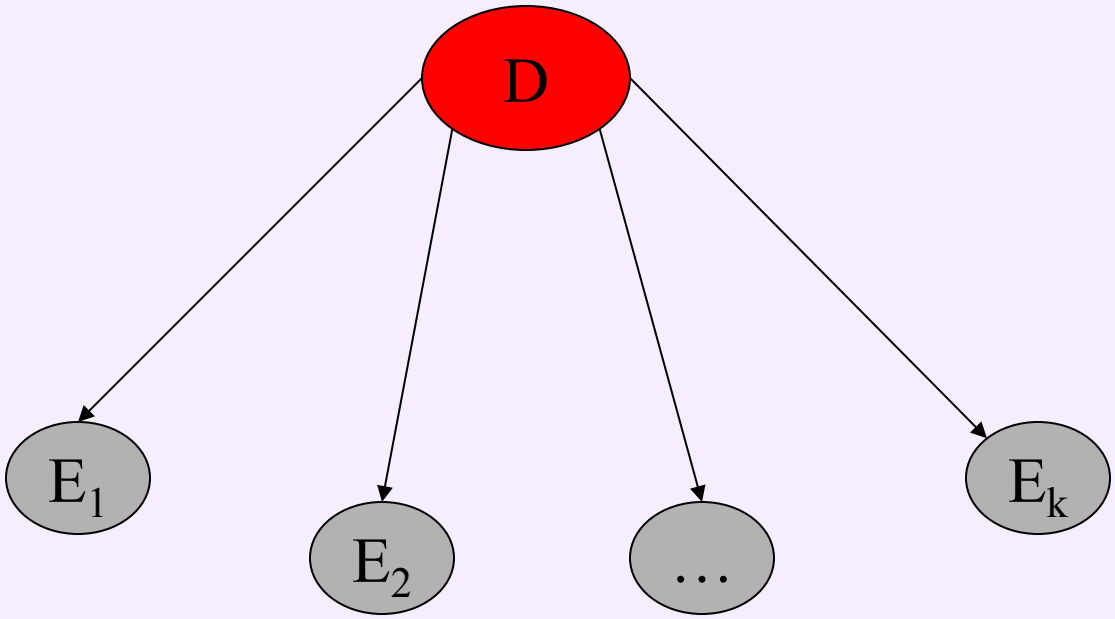
By Theorem 2, the product of these conditional distributions yields a joint probability distribution **P'** of the variables, and **(Γ, P')** satisfies the Markov condition (and it is therefore a **BN**). That is, there exists a joint probability distribution but, in general, it is different than P!

What distribution is then P' ? It is our joint subjective probability distribution of the variables, obtained from our beliefs concerning conditional independencies among the variables (reflected by the structure of the DAG Γ, the CPTs and the marginal distribution of root nodes).

If we are correct about our beliefs, we will obtain P'=P !!

5. Diagnosis and prognosis with BN

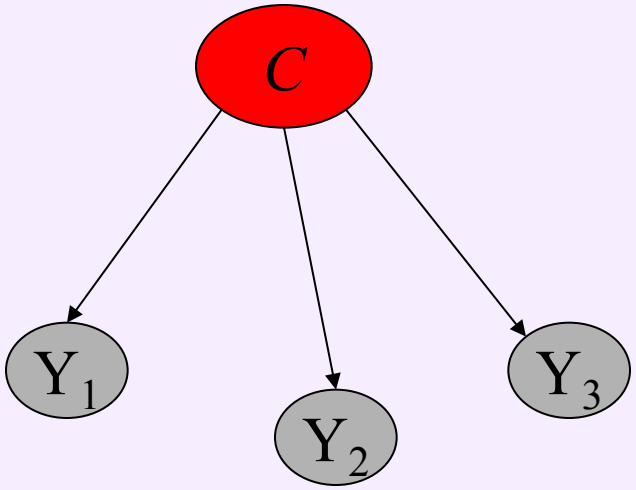
For example, medical **diagnosis** involves building a model for each disease given the set of observed findings that a patient is suffering from. D=disease, E_i=evidences (symptoms, signs, laboratory test results,…).



A (naïve) Bayes Classifier

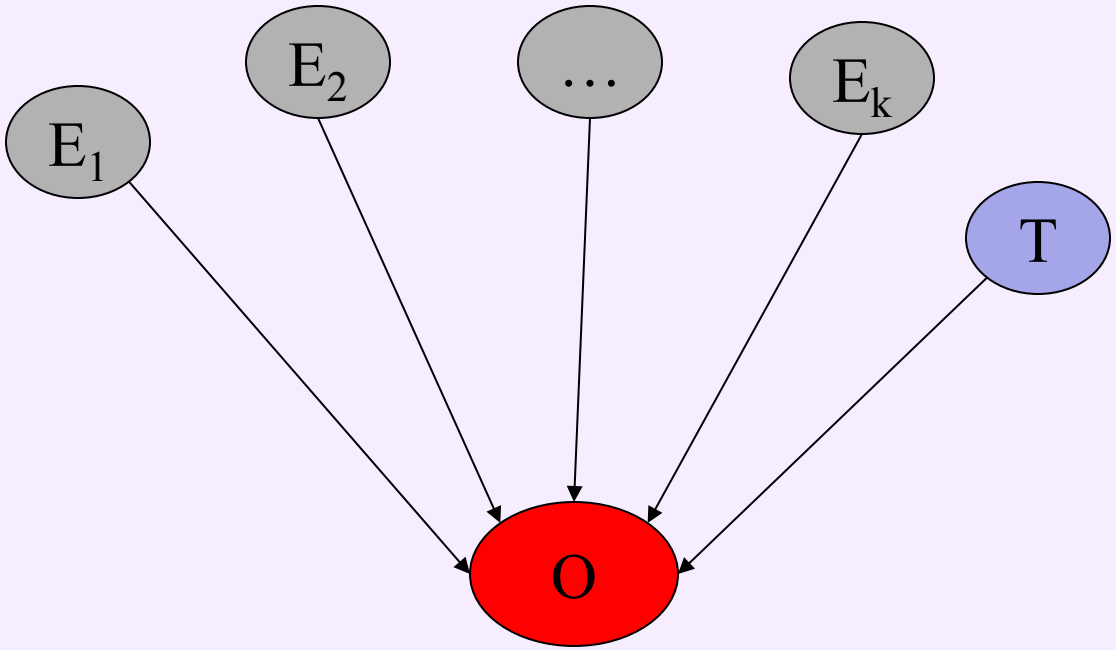
It is a particular case of **diagnosis** with a BN in which the value of a particular feature Y_i is unrelated to the presence or absence of any other feature, given the class variable C.

Example: C=type of fruit, Y_1=color, Y_2=shape, Y_3=size
(a fruit may be considered to be an apple if it is red, round, and about 3" in diameter).
Each of these features contribute independently to the probability that the fruit is an apple, regardless of the presence or absence of the other.



Prognosis with BN

For example, medical **prognosis** attempts to predict the future state of a patient presenting with a set of observed and assigned treatment. O=outcome (life expectancy, quality of life, spread of a disease,…), E_i=evidences (symptoms, signs, laboratory test results,…), T=prescribed treatment.



6. Causality

We study **causality** because we need to estimate the effect of smoking on lung cancer, of education on salaries, of carbon emissions on the climate…

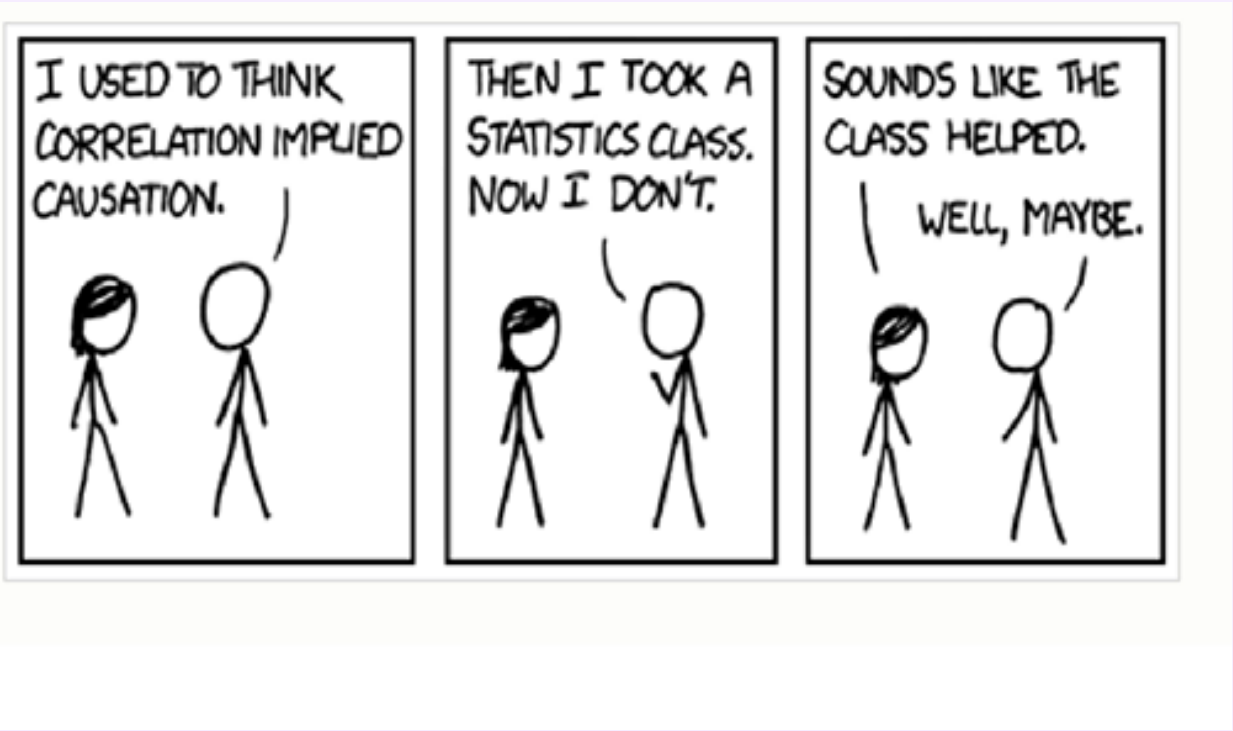
We also need to understand **how** and **why** causes influence their effects: knowing whether malaria is transmitted by mosquitoes or ``mail-air'', as many believed in the past, tells us whether we should pack mosquito nets or breathing masks on our next trip to the swamps.

If the action of making a variable *X* take some value (**manipulate** X) sometimes changes the value taken by variable *Y*, the probability distribution of *Y*, then we assume that *X* is responsible for sometimes changing *Y*'s value, and we conclude that ***X* is a cause of *Y***.

We assume that

causes and their effects are statistically correlated...

But variables can be correlated without one causing the other!



As correlation means not causation, there is no statistical method that can determine the causal story from the data alone. There is, in fact, no way to represent any causal information in contingency tables, on which statistical inference is often based.

There are, however, extra-statistical methods that can be used to express and interpret causal assumptions. One of them is *Bayesian Networks*: they are the focus of this part of the module. With the help of *Bayesian Networks*, you will be able to mathematically describe causal scenarios of any complexity, and answer decision problems.

But *Bayesian Networks* will allow to consider more intricate problems where intuition can no longer guide the analysis.

Example 5

The pharmaceutical company Merck had been marketing its drug **finasteride** as medication for men with benign prostatic hyperplasia.

Based on anecdotal evidence, it seemed there was a correlation between use of the drug and regrowth of scalp hair.

Should Merck conclude that finasteride causes hair regrowth and marked it as a cure for baldness?

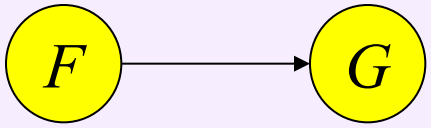
NOT NECESSARILY!

Variables: *F*=finasteride, *G*=growth hair

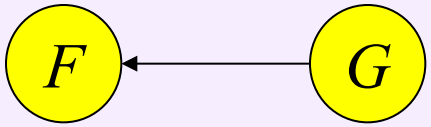
F and *G* can be **not** causally related at all,
or there can be different possible causal relationships
(alone or in combination):

(⟶ denotes causal influence)

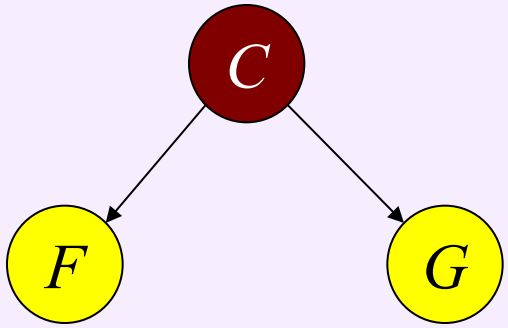
a) *F* causes *G*



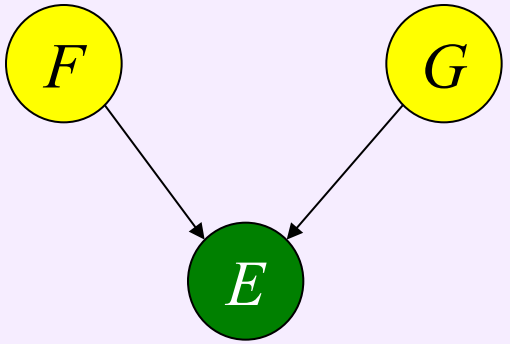
b) *G* causes *F*

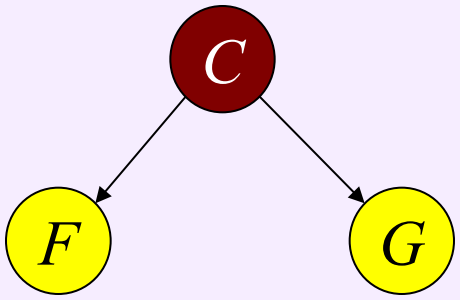


c) Hidden common cause



d) Common effect





c) Hidden common cause

A man concerned about hair loss might try both finasteride and minoxidil in his effort to regrow hair. The minoxidil might cause hair regrowth, whereas the finasteride might no.

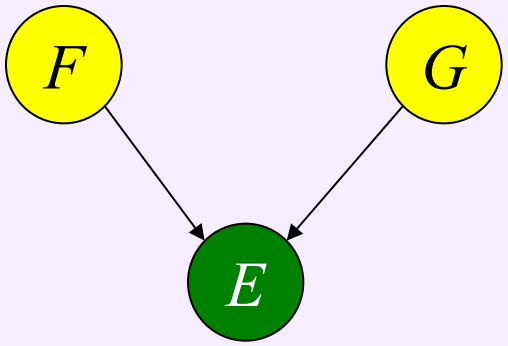
If *C*=the man's concern, *C* is a cause of both finasteride use and hair regrowth (through minoxidil use), whereas the two are not causally related.

Why two variables with a common cause would be correlated?

If *C* is a common cause of *F* and *G,* and neither *F* nor *G* cause the other, suppose c is a value of *C* that has causal influence on *F* taking value f and *G* taking value g.

Then, if *F* had value f, each of its causes would become more probable because one of them should be responsible, so **P(*C*=c/*F*=f)>P(*C*=c)**.

Now, since the probability of c has gone up, that of g would also go up since c causes g. Then, **P(*G*=g/*F*=f)>P(*G*=g)**, which means *F* and *G* correlated.



d) Common effect

Imagine (only for the sake of illustration!) that finasteride and apprehension about lack of hair regrowth both cause hypertension *E*.

Imagine that all the individuals of our sample study have hypertension. We say *E* has been **instantiated**.

This creates dependency between *F* and *G* and would explain the correlation between *F* and *G*.

Why? Because each cause explains the occurrence of the effect, thereby making the other cause less likely.

This type of dependency is called **selection bias**.

Manipulation study in example 5: a randomized controlled experiment (RCE)

Since Merck could not conclude that finasteride causes hair regrowth from their mere correlation alone, they did a manipulation study to test this conjecture.

Study: about 1879 men aged 18 to 41 with mild to moderate hair loss. Half of the men took 1mg. finasteride; the other half, 1mg. of placebo.

Variables:

G (g1=hair regrowth, g2=no)

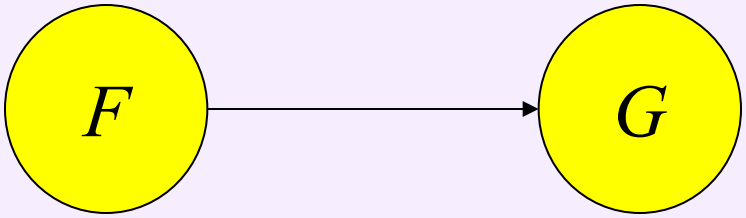
F (f1=subject takes finasteride, f2=subject takes placebo)

Independent dermatologists evaluated photos of the men after 24 months of treatment and found significant regrowth in

- **67%** of the men treated with finasteride,
- **7%** of men treated with placebo.

The RCE supports the conjecture that F causes G to the extent that the data support

$$P(G=g1 \,/\, F=f1) > P(G=g1)$$



F	P(F)
f1	0.5
f2	0.5

Known from the structure of RCE.

F	G	P(G/F)
f1	g1	0.67
f1	g2	0.33
f2	g1	0.07
f2	g2	0.93

Estimated from experiment

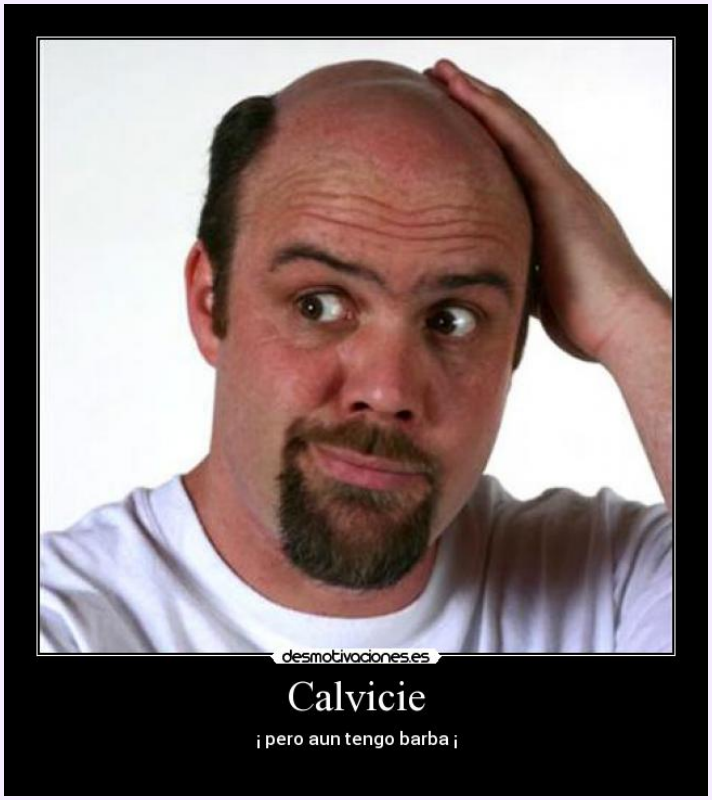
Conclusion: as

$$P(G=g1) = P(G=g1 / F=f1)\, P(F=f1) + P(G=g1 / F=f2)\, P(F=f2)$$
$$= 0.67 \times 0.5 + 0.07 \times 0.5 = 0.37,$$

and then

$$\mathbf{P(G=g1 / F=f1)=0.67 > P(G=g1)=0.37}$$

it can be said that RCE supports the conjecture.



Consequences: Merck concluded that **finasteride** does indeed cause hair regrowth and on December 22, 1997, announced that the U.S.Food and Drug Administration granted marketing clearance to **Propecia™** (finistaride 1mg.) for treatment of male pattern hair loss (androgenetic alopecia), for use in men only.

7. Causal networks

DEFINITION:
We say that ***X* is a cause of *Y*** if a manipulation of *X* results in a change in the probability distribution of *Y*.

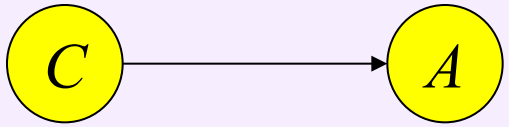
A **causal graph** is a directed graph containing a set of causally related r.v. **V** such that for every *X*, *Y* in **V**, there is an edge from *X* to *Y* if and only if *X* is a cause of *Y*, and there is no subset of variables $\mathbf{W}_{XY}$ of **V** such that if we knew the values of the variables in $\mathbf{W}_{XY}$, a manipulation of *X* would no longer change the probability distribution of *Y*.

If there is an edge from *X* to *Y*, we call ***X* a direct cause of *Y*** (whether or not *X* is a direct cause of *Y* depends on the variables included in **V**).

A causal graph is a **causal DAG** if the graph is acyclic (i.e., there are no causal feedback loops).

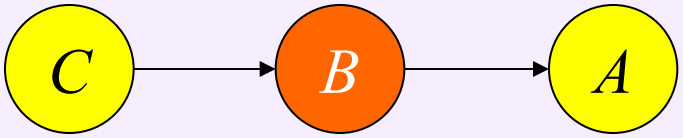
Example 6

If C = "striking a match", A = "the match catching on fire", and no other events are considered, then C is a direct cause of A.



If, however, we added B = "the sulphur on the match tip achieved sufficient heat to combine with oxygen", then we could no longer say that C directly caused A, but rather C directly caused B and B directly caused A.

We say that *B is a causal mediary between C and A* if C causes B and B causes A.



Clearly, we can add more causal *mediaries*. Therefore, rather than assuming that there is a set of causally related variables out there, we will only assume that, in a given context, we identify certain variables and develop a set of causal relationships among them.

DEFINITION:

Let Γ be a **causal DAG** containing the variables of a set of r.v. V.
Let P be the observed probability distribution of V.

Then, if we assume that (Γ, P) satisfies the Markov condition, we say that we are making **the causal Markov assumption**, and we call

(Γ, P) a **causal network**.

Why should we make the causal Markov assumption?

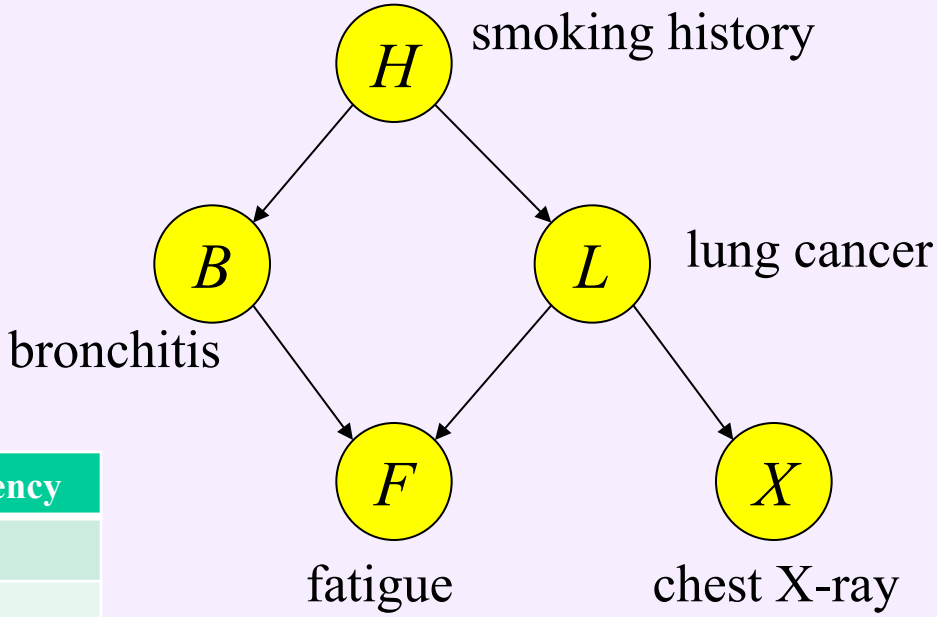
Example 7

A history of smoking is know to cause both bronchitis and lung cancer. Bronchitis and lung cancer both cause fatigue, but only lung cancer can cause a chest X-ray to be positive. No other causal relations.

V=$\{H,B,L,F,X\}$ **E**=$\{(H,B), (H,L), (B,F), (L,F), (L,X)\}$

Γ=(V, E) is a causal DAG

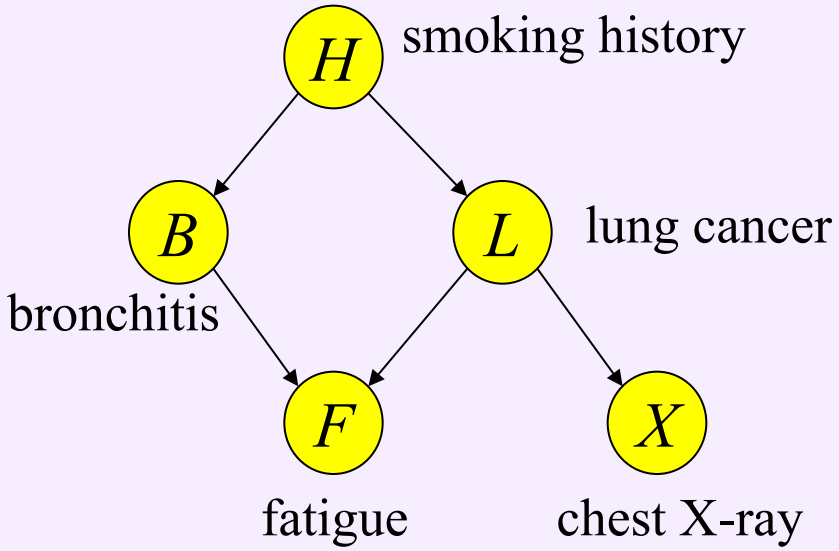
Represents the identified causal relationships among the variables.



The Markov condition:

Node	Parents	Non-desc.	Conditional Independency
H	Null set	Null set	None
B	*H*	*L, X*	*B* and {*L, X*}, given *H*
L	*H*	*B*	*L* and *B*, given *H*
F	*B, L*	*H, X*	*F* and {H, *X*}, given {*B, L*}
X	*L*	*H, B, F*	*X* and {*H, B, F*}, given *L*

$$I_p(\{X\}, ND(X) - PA(X) / PA(X))$$



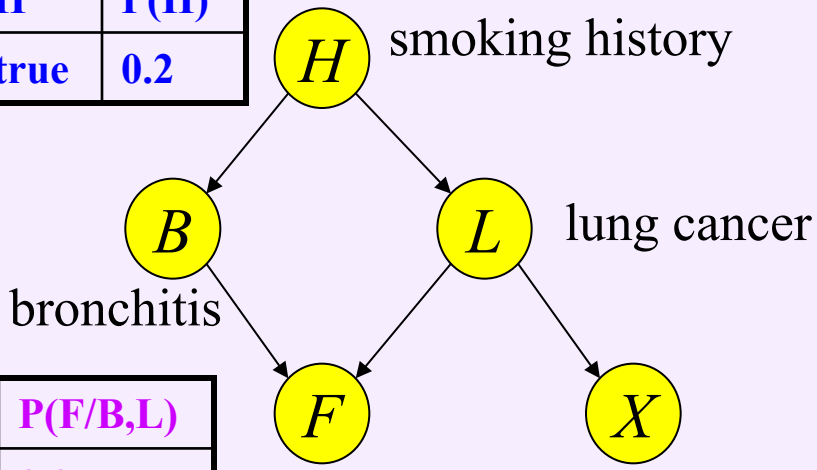
Node	Parents	Non-desc.	Conditional Independency
H	Null set	Null set	None
B	*H*	*L, X*	*B* and {*L, X*}, given *H*
L	*H*	*B*	*L* and *B*, given *H*
F	*B, L*	*H, X*	*F* and {H, *X*}, given {*B, L*}
X	*L*	*H, B, F*	*X* and {*H, B, F*}, given *L*

So, it seems that the Markov condition must hold.
With which probability distribution P?

Informally: we would not expect *B* and *L* to be independent, since if someone had lung cancer it would make it more probable he/she smoked, which would make more probably to have bronchitis.

But if we know he/she smoked, it would already be more probable he/she had bronchitis, and learning that had lung cancer could no longer increase the prob. of smoking (=1), and then, nor that of bronchitis. That is, *B* and *L* are independent given *H*.

Analogously for the other cases.



H	P(H)
true	0.2

H	B	P(B/H)
false	true	0.05
true	true	0.25

H	L	P(L/H)
false	true	0.00005
true	true	0.003

B	L	F	P(F/B,L)
false	false	true	0.05
true	false	true	0.10
false	true	true	0.5
true	true	true	0.75

L	X	P(X/L)
false	true	0.02
true	true	0.6

Let consider P be the product of these conditional distributions, which is a joint probability distribution of the variables **V**, and (Γ, P) satisfies the Markov condition by Theorem 2. So, indeed,

(Γ, P) is a causal network

The causal Markov assumption is justified for a **causal graph** if the following conditions are satisfied:

1. There are no hidden common causes. That is, all common causes are represented in the graph (see Example 8 in what follows).

2. There are no causal feedback loops. That is, our graph is a DAG.

3. Selection bias is **not** present (see Example 9 in what follows).

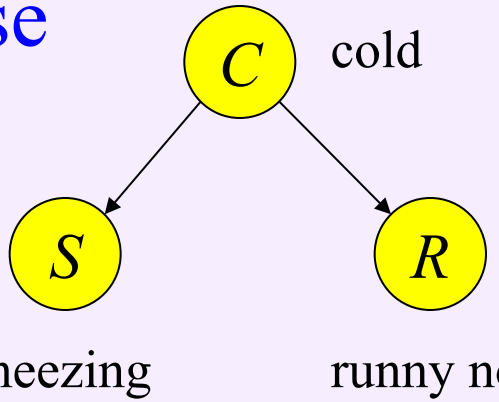
Recall that in a causal graph, there is an edge from X to Y if and only if X is a direct cause of Y.

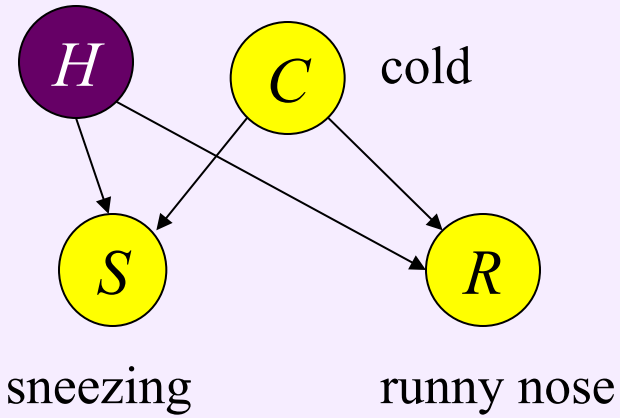
The condition that is most frequently violated is 1.

A DAG can satisfy the Markov condition with the probability distribution of the variables in the DAG without the edges being causal (see Example 1).

Example 8: a hidden common cause

A cold can cause both sneezing and a runny nose, and neither of these conditions can cause each other. Then,we can create the causal DAG:





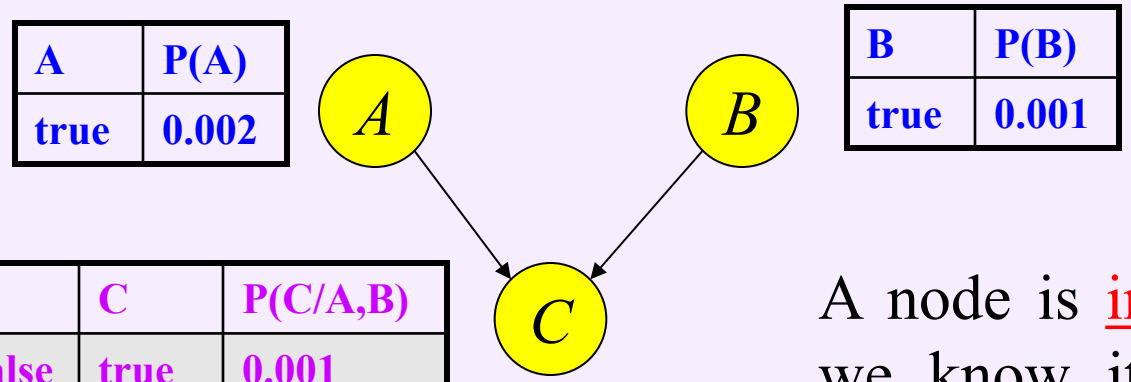
The Markov condition for that DAG would entail that *S* and *R* are independent conditionally to *C*.

However, if there were a hidden common cause of *S* and *R*, *H*, this conditional independency would not hold because even if the value of *C* were known, *S* would change the probability of *H*, which in turn would change that of *R*.

For instance, take ***H*=hay fever**.

The Markov condition fails!

Example 9: selection bias



A	P(A)
true	0.002

B	P(B)
true	0.001

A	B	C	P(C/A,B)
false	false	true	0.001
false	true	true	0.29
true	false	true	0.94
true	true	true	0.95

A node is instantiated when we know its value for the situation being modelled.

Ordinarily, the instantiation of a common effect *C* creates dependency between its causes *A* and *B* because each cause explains away the occurrence of the effect, thereby making the other cause less likely.
This fact would explain correlation between causes A and B.
Recall this dependency is called **selection bias**.

✓ Exercise: show that although *A* and *B* are **independent** (why?), if we instantiate *C*, then they become dependent.

8. References

- "Learning Bayesian networks" by R.E. Neapolitan, Prentice Hall Series in Artificial Intelligence, 2004.
- "Probabilistic Methods for Bioinformatics with and Introduction to Bayesian Networks" by R.E. Neapolitan, Elsevier, 2009.
- "Modeling and reasoning with Bayesian Networks" by A. Darwiche, Cambridge University Press, 2009.
- "Causal Inference in Statistics. A primer" by J. Pearl, M. Glymour and N.P.Jewell, Wiley, 2016.
- "Bayesian networks without tears" by Eugene Charniak.
- "Bayesian Networks. A practical guide to applications" by O. Pourret, P. Naïm and B. Marcot, Wiley 2008.
- "Bayesian Networks and Probabilistic Inference in Forensic Science" by F. Taroni, C. Aitken, P. Garbolino and A. Biedermann, Wiley, 2006.
- "Bayesian Networks in R with applications in Systems Biology" by R. Nagarajan, M. Scutari and S. Lèbre, Springer, 2013.
- "Bayesian Artificial Intelligence" by K. B. Korb and A. E. Nicholson, 2nd edition, CRC Press, 2011.